DOCUMENT RESUME

ED 177 199 TM 009 693

AUTHOR Loyd, Brenda H.; Hoover, H. D.
TITLE A Comparison of Methods of Vertical Equating.
PUB DATE 9 Apr 79
NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Statistics; Elementary Education; *Equated Scores; Mathematics; Raw Scores; Statistical Analysis; Test Interpretation
IDENTIFIERS *One Parameter Model; *Vertical Equating

ABSTRACT

Rasch model vertical equating procedures were applied to three mathematics computation tests for grades six, seven, and eight. Each level of the test was composed of 45 items in three sets of 15 items, arranged in such a way that tests for adjacent grades had two sets (30 items) in common, and the sixth and eighth grades had 15 items in common. In the experiment sixth grade children took either the sixth or seventh grade tests, while children in seventh grade took sixth, seventh, or eighth grade tests and eighth grade children took seventh or eighth grade tests except for a few (91) who took the sixth grade test. The total group for all three grades was 1,956 pupils. The LOGIST program was used with a one parameter option. Several sets of estimations were obtained for each level for separate groups of children, and the obtained ability estimates were placed on a common scale through the use of overlapping items. Inconsistencies among equatings suggested that the test data did not fit the model of the one parameter logistic curve. Analysis of the content showed differences between tests for each grade level. Finally, equipercentile equating methods were applied to the data with a result that differed from the result of the Rasch method. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED177199

A Comparison of Methods of Vertical Equating

Brenda H. Loyd
University of Iowa

H. D. Hoover
University of Iowa

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Brenda H. Loyd

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Paper presented at

Annual Meeting of

the National Council on Measurement in Education

San Francisco, California

April 9, 1979

TM009 693

A Comparison of Methods of Vertical Equating

Introduction

Standardized achievement test batteries are typically constructed to measure the performance of pupils over a range of several grades. To measure such a wide range of ability most efficiently and accurately, levels are constructed within each subtest area, with each level containing items appropriate in difficulty and content for typical students within a particular grade. In order to provide a developmental score scale which is continuous across all levels it is important to have a technique for determining the relationship between a raw score on one level of the test and a raw score on another level.

This relationship between scores is determined through what is frequently referred to as vertical equating. Horizontal equating (Slinde and Linn, 1977) involves equating scores between two tests which are similar in content and difficulty and intended for the same population of examinees; vertical equating, on the other hand, is an attempt to equate scores on two tests which are intentionally designed to be different in difficulty and intended for groups with different ranges of abilities, but measure the same general area of knowledge or general domain of skills.

One method which has been applied to the problem of vertical equating is equipercentile equating. One procedure for equipercentile equating involves the use of two groups of comparable ability. Each group is administered a different form or a different level of the test. The scores on the two are then equated by finding the raw scores on each level of the test which correspond to the same percentile rank on the

other form. In equipercentile equating, however, the equivalence of scores is dependent upon the ability and other characteristics of the population from which the groups are drawn; therefore the conversion scale for any given group may not be altogether accurate for other groups. Other disadvantages of equipercentile equating are discussed by Angoff (1971) and by Slinde and Linn (1977).

In response to the difficulty created by sample dependency in equating and calibrating forms and levels of tests, another method, utilizing the Rasch model approach to item analysis, has received consideration. Latent trait models such as the Rasch model have the proposed advantage of providing item parameters which allow the equating of levels or forms to be independent of the sample used in obtaining the estimates of item parameters. If certain conditions are met the models also produce 'equal interval' scaling along the ability continuun, the underlying latent trait dimension. The Rasch model makes no assumptions about the distribution of ability or the item parameters, and thus has been described as person-free item calibration or test-free measurement.

The Rasch, or one parameter logistic, model specifies an item characteristic curve which defines the relationship between an examinee's ability measured by the items of a test and the probability of answering an item correctly. This item characteristic curve is represented by the following mathematical expression:

$$P_g(\theta) = \frac{e^{Da\ (\theta - b_g)}}{1 + e^{Da\ (\theta - b_g)}}$$

where: $P_g(\theta)$ = the probability of an examinee with ability θ answering item g correctly,

b_g = the item difficulty parameter,

a = the item discrimination parameter, and

D = the scaling constant.

Several conditions must be met for this model to be appropriately used for vertical equating. One condition is the unidimensionality of the ability or trait being measured. Among other conditions (Hambleton and Cook, 1977) are uniform discrimination for all items and the practical elimination of guessing. In addition, for the Rasch model the product of the discrimination parameter and the scaling constant is set equal to one.

Estimates of item characteristic curve parameters can be obtained through the use of computer programs such as those developed by Wood, Wingersky and Lord (1976) and Wright and Mead (1976). The LOGIST computer program (Wood, Wingersky and Lord, 1976) employs a maximum likelihood method to obtain simultaneous estimates of item difficulty parameters and ability parameters for each examinee. The scaling constant is 1.7 and options can be specified in the program to meet the assumptions of the one parameter model, so that the estimated item discrimination, a, is constant across items.

Ability estimates obtained separately for two groups may then be equated using the item difficulty estimates of the common items and the estimates of item discrimination for each set of items with procedures described by Marco (1977).

Under the one-parameter (Rasch) model, a one-to-one relationship is assumed between the raw score (estimated true score) and the estimated ability of an examinee. Based on this relationship, raw scores on two forms or levels can be equated by determining the ability estimate that corresponds to a particular raw score on one form or level, and then equating that raw score with the raw score on the other form or level that has the corresponding ability estimate.

Slinde and Linn (1978) utilized the Rasch model to equate subsets of items of a college math achievement test. The items of the entire test were divided into two subsets, one containing difficult items and one containing easy items. The examinees were divided into three groups (high, middle and low ability) on the basis of their scores on the easy subset of items. Separate calibration of ability estimates on the difficult and easy subsets indicated that an examinee of middle ability would do better to take a more difficult test if the ability estimates were obtained from a high ability group, and would do better to take an easier test if the ability estimates were obtained from a low ability group. Slinde and Linn did not consider the equating with the Rasch model satisfactory, but they suggested that the subsets used represented a more extreme difference in difficulty than would normally be found in the vertical equating of levels of a test.

Another study (Guskey and Katims, 1978) utilized the Rasch model for equating levels of a reading test composed of six levels of difficulty, with adjacent levels containing overlapping items. The overlapping items were used to determine the relationship among ability estimates from the separate levels, and all six levels were then placed on a common ability scale. The results of this vertical equating of the levels of the reading test were reported to be stable and consistent across all levels.

The purpose of the present study is to explore a practical application of the Rasch model to the vertical equating of levels of a mathematics computation test. Three aspects of the problem will be considered: 1) the consistency of calibration of ability estimates on the same test level for different levels of ability; 2) the adequacy of

vertically equating adjacent test levels when estimates of parameters are obtained from two groups of comparable ability; and 3) comparison of the results of equating directly two nonadjacent test levels with the results of equating them by pairwise chaining through an intermediate level. If the Rasch model is appropriate for vertical equating of this test, the calibrations should be consistent in determining ability estimates for the separate ability groups, and the equating of both adjacent and nonadjacent levels should be invariant with respect to groups.

Procedure

The achievement test used in this study is a mathematics computation test composed of three overlapping levels, Levels 12-14, which were built to be appropriate in difficulty and content for typical students of grades 6-8, respectively. All items in the test were tried out and selected for content, appropriate difficulty, and discrimination within each grade. Each level is composed of 45 items; adjacent levels (12 and 13, 13 and 14) have 30 items in common, while nonadjacent levels (12 and 14) have 15 items in common. This overlapping of items is demonstrated graphically in Figure 1.

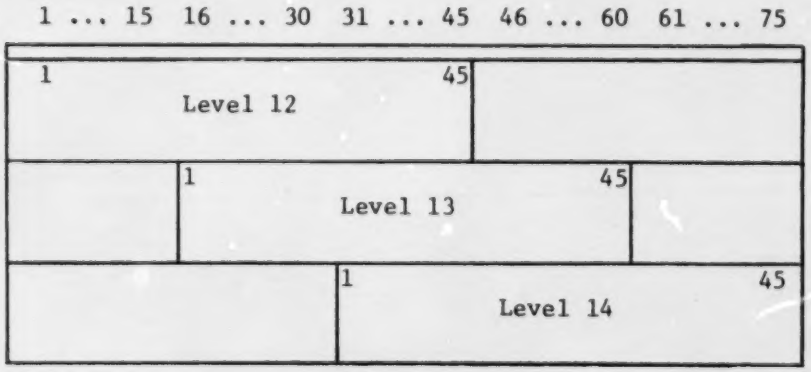

Figure 1.

The 1,956 pupils selected to participate in the equating project were a representative sample of sixth, seventh and eighth grade students in the state of Iowa. Within each classroom (a classroom contained students of one specific grade) pupils were randomly assigned a level of the mathematics computation test such that a proportionate number of each level of interest was administered to relatively comparable ability groups of pupils within each grade. Sixth grade pupils were administered either Level 12 (N = 259), the on-level test, or Level 13 (N = 252), an out-of-level test (one level above usual level for sixth graders). Seventh grade pupils were administered either Level 13 (N = 253), the on-level test, or Level 14 (N = 257), an out-of-level test (one level above the usual level for seventh graders), or Level 12 (N = 245), an out-of-level test (one level below the usual level for seventh graders). Eighth grade pupils were administered either Level 14 (N = 301), the on-level test (i.e., the usual test for eighth graders), or Level 13 (N = 298), an out-of-level test (one level below the usual level for eighth graders); eighth grade pupils were also administered Level 12 (N = 91), the out-of-level test two levels below the usual level for eighth graders. The assignment of these tests is shown graphically in Figure 2.

	Level 12	Level 13	Level 14
Grade 6	(N=259)	(N=252)	
Grade 7	(N=245)	(N=253)	(N=257)
Grade 8	(N= 91)	(N=298)	(N=301)

Figure 2.

Estimates of item difficulty parameters and estimates of ability were obtained with the aid of the LOGIST computer program, utilizing the options which permit estimates under the one parameter logistic model. These estimates differ from those obtained under the Rasch model only in terms of unit of scale. Several sets of estimations were obtained for each level for separate groups of pupils. Estimates of parameters for items on Level 14 were calculated separately for pupils of grade 7 and grade 8. Likewise, estimates for items on Level 13 were calculated separately for pupils of grade 6, grade 7 and grade 8. Finally, estimates for items on Level 12 were calculated separately for pupils of grade 6, grade 7 and grade 8. To determine the relationship among the separate estimations, the obtained ability estimates were placed on common scales through the use of the overlapping items. Ability estimates obtained on the same level test (i.e., for a single level), have 45 items in common, 30 items were in common for adjacent levels, and for nonadjacent levels 15 items were in common.

Raw scores on each level for each group of pupils were equated by determining the ability estimate that was calculated to correspond to a particular raw score, and then determining the raw score that corresponded to the same ability estimate obtained in a separate calculation. Raw scores were thus equated for 1) the same level equated to itself through different ability groups (different sets of pupils), Level 12, Level 13 and Level 14 each equated to itself; 2) Level 12 equated to Level 13 and Level 13 to Level 14 through calibrations based on comparable ability groups, e.g., Level 12-Grade 6 raw scores equated to Level 13-Grade 6 raw scores; 3) Level 12 equated to raw scores on Level 14, directly through comparable ability groups; and 4) Level 12 equated to raw scores on Level 14 after first being equated to Level 13. The levels and groups equated are presented in Table 1.

Table 1

Test Levels and Ability Groups Used in Equatings

		Base Level	Calibration Grade Group	Equated Level	Calibration Grade Group	Items Used in Calibration	Equating Method Code
Same Level Equated		12	6	12	7		S12-6,7
		12	6	12	8		S12-6,8
		13	7	13	6	45	S13-7,6
		13	7	13	8		S13-7,8
		14	8	14	7		S14-8,7
Adjacent Levels Equated		14	7	13	7		A14,13-7
		14	8	13	8		A14,13-8
		13	6	12	6	30	A13,12-6
		13	7	12	7		A13,12-7
		13	8	12	8		A13,12-8
Non-Adjacent Levels Equated (14-12)	Direct	14	7	12	7	15	D-7
		14	8	12	8		D-8
	Chaining	14	7	12	7		C-7
		14	8	12	8	30	C-8
		14	8	12	6		*C-6,7,8

*For this equating both test level and calibration group were allowed to vary. Level 14 was equated to Level 13 using 8th graders for Level 14 and 7th graders for Level 13. Likewise, Level 13 was equated to Level 12 using 7th graders for Level 13 and 6th graders for Level 12. That is, all calibrations were based on 'on-level' students.

Results and Discussion

Results of equating raw scores under the one parameter logistic model are presented in Tables 2, 3 and 4. The equating of Level 14 to Level 13 with the 30 items in common (Table 2) produced different results for comparable groups of seventh graders and comparable groups of eighth graders. The mean difference between raw scores (for Level 13 raw scores of 10 to 40) on Level 14 for the two equating groups was 1.43 raw score points and the maximum difference between the two equatings was approximately 1.8 raw score points. Equatings of Level 12 to Level 13 also involved inconsistencies (Table 2). Level 12 was equated to Level 13 through three grade groups, sixth, seventh and eighth. The mean difference between the sixth grade and seventh grade equatings on the Level 12 raw score scale was 1.48 raw score points. The mean difference for the seventh grade and eighth grade equatings was 1.22 raw score points.

The result of directly equating Level 14 to Level 12 involving 15 common items and the indirect equating of Level 14 to Level 12 involving pairwise chaining through Level 13 with common sets of 30 items are presented in Table 3. The degree of consistency between any two equatings varies with the method used to equate and the groups involved in the calibration. The mean difference between the two direct equatings, one with seventh grade pupils and the other with eighth grade pupils, was 2.09 raw score points where the differences were determined between raw scores of 11 and 36 on Level 14.

Table 4 contains the results of equating a level to itself (45 items in common) with different grade groups. Level 14 equated with seventh and eighth grade groups resulted in a mean difference of .31 raw score points between raw scores of 10 and 40 on Level 14. The mean

Table 2

'Equivalent' Raw Scores for Adjacent Levels

Level 13 Raw Score	Level 14 Raw Score	
Equating Method	A14,13-7	A14,13-8
1		
2		
3		
4		
5		
6		
7		
8		
9		4.9
10	4.9	5.5
11	5.5	6.1
12	6.1	6.8
13	6.7	7.6
14	7.4	8.3
15	8.0	9.0
16	8.7	9.8
17	9.4	10.6
18	10.2	11.4
19	10.9	12.2
20	11.6	13.0
21	12.4	13.9
22	13.2	14.7
23	14.1	15.7
24	14.9	16.6
25	15.9	17.5
26	16.8	18.5
27	17.8	19.4
28	18.8	20.4
29	19.8	21.4
30	20.8	22.5
31	21.8	23.6
32	23.0	24.8
33	24.2	26.0
34	25.4	27.2
35	26.8	28.5
36	28.0	29.8
37	29.3	31.0
38	30.6	32.4
39	32.2	33.8
40	33.7	35.2
41	35.3	36.8
42	37.0	38.4
43	39.0	40.2
44		42.3
45		

Level 13 Raw Score	Level 12 Raw Score		
Equating Method	A13,12-6	A13,12-7	A13,12-8
1			
2			
3	9.0		
4	11.3		
5	13.2	11.0	10.9
6	14.8	12.7	12.5
7	16.4	14.3	13.9
8	18.0	15.9	15.3
9	19.4	17.4	16.6
10	20.6	18.8	17.9
11	23.0	20.0	19.1
12	24.2	21.4	20.3
13	25.4	22.6	21.3
14	26.3	23.8	22.5
15	27.4	25.0	23.6
16	28.4	26.0	24.7
17	29.4	27.2	25.9
18	30.2	28.3	26.9
19	31.0	29.4	28.0
20	31.9	30.4	28.9
21	32.8	31.2	29.8
22	33.6	32.1	30.7
23	34.5	33.1	31.6
24	35.2	33.9	32.4
25	36.1	34.8	33.2
26	36.8	35.6	33.9
27	37.5	36.3	34.8
28	38.1	37.0	35.6
29	38.8	37.8	36.4
30	39.4	38.4	37.1
31	40.0	39.0	37.8
32	40.6	39.7	38.5
33	41.2	40.2	39.2
34	41.7	40.9	39.8
35	42.1	41.4	40.4
36	42.6	41.9	41.0
37	43.0	42.4	41.5
38	43.4	42.8	42.0
39	43.8	43.2	42.6
40		43.8	43.0
41		43.9	43.5
42			43.9
43			
44			
45			

Table 3

'Equivalent' Raw Scores for Nonadjacent Levels

Level 14 Raw Score	Level 12 Raw Scores				
Equating Method	D-7	D-8	C-7	C-8	C-6,7,8
1					
2					
3					
4					
5	17.9		19.0		23.2
6	20.0		21.2	16.6	24.8
7	21.9		23.3	19.2	26.2
8	23.6		24.9	21.4	27.4
9	25.3		26.4	23.2	28.5
10	27.0		28.0	24.7	29.4
11	28.4	25.4	29.2	26.2	30.5
12	29.7	26.8	30.6	27.5	31.4
13	30.8	27.9	31.8	28.8	32.2
14	31.9	28.9	32.8	29.8	32.9
15	32.9	30.1	33.9	30.9	33.6
16	33.8	31.0	34.8	31.9	34.4
17	34.7	32.0	35.6	32.7	35.0
18	35.7	32.9	36.5	33.6	35.6
19	36.3	33.8	37.1	34.4	36.2
20	37.0	34.6	37.9	35.2	36.9
21	37.8	35.5	38.5	36.0	37.4
22	38.4	36.1	39.1	36.7	38.0
23	39.0	36.8	39.6	37.4	38.5
24	39.6	37.4	40.1	38.0	38.9
25	40.2	38.1	40.6	38.6	39.4
26	40.7	38.7	41.1	39.1	39.8
27	41.1	39.3	41.5	39.6	40.2
28	41.6	39.8	41.9	40.2	40.6
29	41.9	40.3	42.3	40.6	41.0
30	42.3	40.7	42.6	41.0	41.4
31	42.6	41.2	42.9	41.5	41.7
32	42.9	41.7	43.2	41.9	42.1
33	43.2	42.0	43.4	42.3	42.4
34	43.5	42.4	43.7	42.7	42.7
35	43.7	42.8	43.9	43.0	43.0
36	43.9	43.1	44.1	43.3	43.3
37		43.4		43.6	43.6
38		43.7		43.8	43.8
39		43.9		44.0	44.1
40					
41					
42					
43					
44					
45					

Table 4

Equating of Each of Levels 12-14 to Itself

Level 14	
Equating Method	S14-8,7
1	
2	
3	
4	
5	
6	6.4
7	7.8
8	9.0
9	10.1
10	11.2
11	12.2
12	13.1
13	14.0
14	14.9
15	15.7
16	16.6
17	17.5
18	18.4
19	19.3
20	20.1
21	21.0
22	22.0
23	22.8
24	23.6
25	24.6
26	25.4
27	26.3
28	27.2
29	28.0
30	28.9
31	29.8
32	30.8
33	31.7
34	32.7
35	33.7
36	34.7
37	35.7
38	36.7
39	37.8
40	38.8
41	39.9
42	40.9
43	
44	
45	

Level 13		
Equating Method	S13-7,6	S13-7,8
1		
2		
3		
4		
5		
6	7.0	4.8
7	8.1	5.7
8	9.2	6.8
9	10.2	7.7
10	11.2	8.7
11	12.2	9.8
12	13.2	10.8
13	14.2	11.9
14	15.1	13.0
15	16.1	14.1
16	17.0	15.2
17	17.9	16.3
18	18.8	17.4
19	19.8	18.5
20	20.7	19.7
21	21.5	20.8
22	22.4	22.0
23	23.3	23.1
24	24.2	24.3
25	25.0	25.5
26	25.9	26.6
27	26.8	27.8
28	27.7	29.0
29	28.6	30.0
30	29.5	31.2
31	30.4	32.2
32	31.4	33.3
33	32.2	34.3
34	33.1	35.4
35	34.0	36.2
36	35.0	37.2
37	36.0	38.3
38	36.9	39.2
39	38.0	40.2
40	39.0	41.1
41	39.9	41.9
42		42.7
43		43.5
44		
45		

Level 14		
Equating Method	S12-6,7	S12-6,8
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		8.3
11	9.6	9.2
12	10.6	10.2
13	11.6	11.2
14	12.7	12.2
15	13.8	13.3
16	14.8	14.3
17	15.9	15.3
18	17.0	16.4
19	18.1	17.6
20	19.2	18.7
21	20.3	19.8
22	21.4	20.9
23	22.6	22.2
24	23.7	23.2
25	24.9	24.5
26	26.0	25.8
27	27.3	27.1
28	28.5	28.2
29	29.7	29.7
30	30.8	31.0
31	32.0	32.2
32	33.1	33.3
33	34.2	34.6
34	35.3	35.7
35	36.5	36.9
36	37.5	38.0
37	38.5	39.0
38	39.5	40.0
39	40.4	40.9
40	41.3	41.8
41	42.2	42.6
42	42.9	43.2
43	43.7	43.9
44		
45		

difference for Level 13 equated with sixth and seventh grade groups was .06 raw score points and for Level 13 equated with seventh and eighth grade groups was .06 raw score points again. For Level 12 the mean difference between equating with sixth and seventh grade pupils was .003 raw score points and between sixth and eighth grade pupils was .18 raw score points.

The inconsistencies among equating, especially among the equating of adjacent and nonadjacent levels, suggest that the test data does not fit the model of the one parameter logistic curve. This lack of fit could be due to violations of one or more of the assumptions. For example, item discrimination is not the same for all items in the levels over all grades. The lack of consistency may also be due to violation of the assumption of unidimensionality which is basic to the model; i.e., the traits underlying mathematical computation may change from level to level. The differences in equatings of adjacent and nonadjacent levels for different grade groups indicate that mathematics performance may be differentially dependent upon school curriculum. Performance of the grade groups may be dependent upon the introduction of specific topics in math computation at certain points in the elementary school curruculum. An analysis of the skills classification of items in each level indicates that there are obvious differences in the number of items in each skills category represented in adjacent levels such as Level 12 and Level 13 and differences in the type of skills included at each level. A comparison of skills classifications of items for Level 12 and Level 14 reveals even greater discrepancies in number and type of skills measured under mathematical computation. The skills classification of items in Levels 12-14 are presented in Table 5.

Table 5

Number of Items Measuring Each Skill
for Levels 12-14

Skill	Level		
	12	13	14
Whole Numbers	28	16	11
Addition	5	3	3
Subtraction	5	3	3
Multiplication	9	5	3
Division	9	5	2
Fractions	13	18	15
Addition	4	5	3
Subtraction	6	6	3
Multiplication	3	4	5
Division		3	4
Decimals	4	11	19
Addition	2	3	4
Subtraction	2	3	4
Multiplication		3	5
Division		2	6
TOTALS	45	45	45

Practical Implications of Results

The major reason for the necessity of vertical equating of levels on an achievement battery is to enable the test publisher to establish a developmental score-scale that may be used in following the progress of pupils as they move through the educational system. There are many types of such scales, e.g., grade-equivalent, age-equivalent, and expanded standard scores. However, the type of developmental scale used is essentially independent of the method used in vertical equating, and differences in equating methods will have the same impact on any developmental scale.

Another benefit of having the various levels of an achievement battery

equated is to enable schools to employ what is commonly referred to as testing 'out-of-level.' In most school systems a substantial number of students are enough retarded in their educational development that the level of test intended for use with the 'typical' student at a given grade level is inappropriate, both in terms of difficulty and test content. Such pupils are frequently given a lower level of the test battery. However, it is still desirable to assign within grade percentile ranks to these pupils in order to determine grade group, class, building, or system averages all pupils within a given grade must have derived scores on a common score-scale.

To illustrate the practical effect of the different methods of equating consider an eighth grade pupil tested 'out-of-level' with Level 12. Once this pupil obtains a given raw score on Level 12 the 'equivalent' raw score on Level 14 will be different depending on the equating method used. Table 6 presents Level 14 raw scores corresponding to raw scores of 20, 25, 30, 35, and 40 on Level 12 for the five different methods of equating Level 12 to Level 14 previously discussed. Level 14 of the test used in this investigation was nationally standardized in the fall of 1977 and national percentile ranks are available. These percentile ranks are reported in Table 6 and the differences are substantial. For example, if an eighth grade pupil was administered Level 12 and obtained a raw score of 35, his/her grade 8 percentile rank would be 35 if the direct method of equating based on grade 7 pupils was used in establishing the Level 14 raw score. If however the direct method based on grade 8 pupils was used the corresponding percentile rank would be 49. While the differences obtained for pupils are quite large, corresponding differences for school averages would be much more pronounced. Distributions of school averages are much more homogeneous than

pupil distributions. The effect of differences in equating of two or three raw score points on averages of developmental scores, and the percentile ranks of these averages, will be much larger than those reported in Table 6. The impact of the different methods is obviously dependent on the proportion of pupils tested 'out-of-level.'

Table 6

The Effect of Different Equating Methods on National Percentile Ranks of Eighth Grade Students Administered Level 12

Equating Method	Direct (15 overlapping items) D-7		D-8		Chaining (30 overlapping items) C-7		C-8		C-6,7,8	
Level 12 RS	Level 14 RS	Grade 8 PR	Level 14 RS	Grade 8 PR	Level 14 RS	Grade 8 PR	Level 14 RS	Grade 8 PR	Level 14 RS	Grade 8 PR
20	6.0	1			5.4	1	7.4	2		
25	8.8	4	10.7	8	8.1	2	10.2	7	6.3	1
30	12.3	13	14.9	24	11.4	10	14.2	21	10.5	8
35	17.3	35	20.3	49	16.2	30	19.2	44	17.0	34
40	24.7	66	28.4	78	23.8	63	27.7	75	26.5	72

The major focus of this paper has been one of investigating differences in equating using the one parameter latent trait model when different calibration groups are used. A number of investigators are studying differences between various types of methods for vertical equating, e.g., latent trait (one, two, and three parameter models), equipercentile, linear, etc. While such comparisons were not the intention of this study one example will be provided.

Probably the most common method of equating used in the past has been equipercentile. Table 7 presents a comparison between the equating method C-7 used in this study and an equipercentile equating using the same pupils. This method was a direct equating of Level 12 to Level 14 where the equating group was 7th grade students.

Table 7

A Comparison of Rasch and Equipercentile Equating of Level 12 to Level 14

Equating Method	Rasch		Equipercentile	
Level 12 RS	Level 14 RS	Grade 8 PR	Level 14 RS	Grade 8 PR
20	6.0	1	12.7	14
25	8.8	4	15.1	25
30	12.3	13	17.7	37
35	17.3	35	21.8	56
40	24.7	66	27.4	75

These differences are even more pronounced than those reported earlier and such differences would obviously have a major impact on derived scores in school systems where a substantial number of students are tested 'out-of-level.' The equipercentile equating is based on all 45 items in both tests and the Rasch equating on only the 15 overlapping items. In two test levels where the content is by necessity quite different such differences were not unexpected.

Summary

While latent trait methods show a great deal of promise for improving the horizontal equating of tests, the results from the present study, and others, indicate that their use in vertical equating should be approached with extreme caution. This warning seems especially appropriate in the vertical equating of achievement tests where by necessity the content specifications of the test appreciably change with test level. A note of caution should also be raised for those test developers assembling achievement test item banks based on item calibration by latent trait methods.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington D. C.: American Council on Education, 1971.

Guskey, T. R. & Katims, M. Use of the Rasch model in calibrating the reading comprehension test of the Iowa Tests of Basic Skills. Department of Research and Evaluation, Chicago Board of Education, March 1978.

Hambleton, R. J. & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 139-160.

Slinde, J. A. & Linn, R. L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Slinde, J. A. & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N. J.: Educational Testing Service.

Wright, B. D. & Mead, R. J. BICAL: Calibrating items and scales with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1976.